METHOD FOR COMPUTING MODELS BASED ON ATTRIBUTES SELECTED BY ENTROPY

RELATED APPLICATIONS

The present invention is related to the subject matter of the following commonly assigned, copending United States patent applications: serial no. 09/282,622 ~~(Docket No. AT9-99-038)~~ entitled "COMPILING MESSAGE CATALOGS FOR CLIENT-SIDE CONSUMPTION" and filed March 31, 1999; serial no. 09/282,620 ~~(Docket No. AT9-99-039)~~ entitled "DYNAMIC SCREEN CONTROL" and filed March 31, 1999; and serial no. 09/282,682 ~~(Docket No. AT9-99-040)~~ entitled "METHOD OF CUSTOMIZING SCREEN AND APPLICATION BEHAVIOR USING ATTRIBUTE METADATA IN AN APPLICATION DATABASE" and filed March 31, 1999. The content of the above-referenced applications is incorporated herein by reference.

BACKGROUND OF THE INVENTION

1. Technical Field:

The present invention relates in general to data analysis and in particular to qualifying sample populations employed in predictive data analysis. Still more particularly, the present invention relates to reducing the number of attributes of a sample population employed in generating a predictive model based on the sample population.

2. Description of the Related Art:

A wide array of subjects are the focus of contemporary data collection, such as customer data which is collected by various industries, medical information which is collected for development of diagnostics and treatment protocols, or data relating to insurable or potentially insurable activities which is collected for insurance risk assessment. Collection of such data has become so routine and pervasive that "data mining" is frequently required to separate useful information from dross.

Data collection is frequently undertaken for the purposes of developing predictive models. That is, by statistical analysis of characteristics of a sample population, attempts are made to derive models which may predict, with reasonable accuracy, whether an individual subject will exhibit a characteristic or group of characteristics of interest based on known characteristics of that subject. For instance, marketing firms may attempt to develop predictive models for determining which individuals within a target population are most likely to respond to a particular promotional campaign.

Contemporary data collection generally proceeds more or less indiscriminately. That is, those engaged in collection of data typically collect as much data regarding each individual subject as possible, without regard to the ultimate usefulness of the data in, for example, developing a predictive model. This may result from uncertainty regarding which characteristics are most useful for a particular purpose and/or the simplistic conviction that

more data will produce better results. More frequently, however, indiscriminate data collection results instead from the use of a data set for more than one purpose, spreading the cost of the data collection among multiple projects.

One effect of indiscriminate data collection on the development of predictive models is the inefficiency and error introduced by large data samples. A sample population may include data for five hundred or more characteristics of each individual subject within the sample population. Attempting to generate a predictive model based on that many individual characteristics is computationally inefficient. Furthermore, as the number of characteristics or attributes employed in generating the predictive model increases, the probability that the sample population is skewed by one or more characteristics or attributes also increases.

It would be desirable, therefore, to provide a mechanism for preprocessing a sample population to reduce the number of attributes or characteristics employed in generating a predictive model. It would further be advantageous if the mechanism eliminated characteristics which might skew the sample population and thereby degrade the accuracy of a predictive model generated from such data.

SUMMARY OF THE INVENTION

It is therefore one object of the present invention to provide improved data analysis.

It is another object of the present invention to provide improved qualification of sample populations employed in predictive data analysis.

It is yet another object of the present invention to provide a technique for reducing the number of attributes of a sample population employed in generating a predictive model based on the sample population.

The foregoing objects are achieved as is now described. Attributes of a data set to be employed in generating a predictive model are analyzed based on entropy, chi-square, or similar statistical measure. A target group of samples exhibiting one or more desired attributes is identified, then remaining attribute values for the target group are compared to corresponding attribute values for the whole sample population. A subset of all available attributes is then selected from those attributes which exhibit, when comparing attribute values of target group samples to attribute values for the whole sample population, the greatest relative difference or divergence. That is, an attribute for which the target group samples exhibit, for example, only two of all possible values is selected in preference to an attribute for which the target group samples exhibit three or more of the possible values. This subset is employed to generate the predictive model. Efficiency in generating the predictive model is improved,

since fewer attributes are employed and less computational resources are required. Accuracy of the resulting predictive model is also improved since attributes potentially skewing the sample population in a manner least related to the desired attribute are eliminated from consideration in developing the model.

The above as well as additional objects, features, and advantages of the present invention will become apparent in the following detailed written description.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features believed characteristic of the invention are set forth in the appended claims. The invention itself however, as well as a preferred mode of use, further objects and advantages thereof, will best be understood by reference to the following detailed description of an illustrative embodiment when read in conjunction with the accompanying drawings, wherein:

Figure 1 depicts a block diagram of a data processing system in which a preferred embodiment of the present invention may be implemented;

Figure 2 is a logical block diagram for a mechanism for preprocessing a sample population to be employed in generating a predictive model in accordance with a preferred embodiment of the present invention; and

Figure 3 depicts a high level flow chart for a process of selecting attributes of a sample for generating a predictive model in accordance with a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

With reference now to the figures, and in particular with reference to **Figure 1**, a block diagram of a data processing system in which a preferred embodiment of the present invention may be implemented is depicted. Data processing system **100** may be, for example, one of the models of personal computers available from International Business Machines Corporation of Armonk, New York. Data processing system **100** includes a processor **102**, which in the exemplary embodiment is connected to a level two (L2) cache **104**, connected in turn to a system bus **106**. In the exemplary embodiment, data processing system **100** includes graphics adapter **116** also connected to system bus **106**, receiving user interface information for display **120**.

Also connected to system bus **106** is system memory **108** and input/output (I/O) bus bridge **110**. I/O bus bridge **110** couples I/O bus **112** to system bus **106**, relaying and/or transforming data transactions from one bus to the other. Peripheral devices such as nonvolatile storage **114**, which may be a hard disk drive, and keyboard/pointing device **116**, which may include a conventional mouse, a trackball, or the like, are connected to I/O bus **112**.

The exemplary embodiment shown in **Figure 1** is provided solely for the purposes of explaining the invention and those skilled in the art will recognize that numerous variations are possible, both in form and function. For instance, data processing system **100** might also include a compact disk read-only memory (CD-ROM) or digital video disk (DVD) drive, a sound card and audio speakers, and numerous

other optional components. All such variations are believed to be within the spirit and scope of the present invention. However, data processing system **100** is preferably programmed to provide a mechanism for preprocessing a sample population to be employed in generating a predictive model by reducing the number of attributes of the sample population which are utilized in generating the predictive model.

Referring to **Figure 2**, a logical block diagram for a mechanism for preprocessing a sample population to be employed in generating a predictive model in accordance with a preferred embodiment of the present invention is illustrated. The mechanism **200** includes a preprocessing module **202** which receives a sample **204**. Sample **204** is preferably a relational database containing a number of data elements **206** (rows). Each data element **206** includes various attributes **208** or characteristics (columns). In the example contemplated, sample **204** may contain any arbitrary number of data elements, and the number of attributes **208** of each data element **206** may equal or exceed 500.

Preprocessing module **202** receives sample **204** in which at least some data elements possess a desired attribute or group of attributes, which may be referred to as a "target" data set. Module **202** then selects other attributes for data elements within sample **204** to be used in generating a predictive model by statistically analyzes sample **204** to determine which attributes of the target data elements differ the most from corresponding attributes of the sample population as a whole. The attributes for which sample instances having a desired characteristic have values which are the most different from corresponding attribute values

of the sample population generally are selected.

For the purposes of this description only, and without intended to imply any limitations to the present invention, sample **204** may be viewed as logically divided within module **202** into a target group **210** and one or more other groups **212** based on the value of a specific attribute. Data elements within sample **204** having the desired attribute value or values are categorized in target group **210**; data elements within sample **204** not having the desired attribute or attributes are categorized in one of the other groups **212**. The other attributes--those not forming the basis of logical division of sample **204**--of target group **210** are then compared to the attributes of entire sample **204** to determine which attributes exhibit the largest difference between target group **210** and sample **204**.

For example, suppose an attribute A may have three possible values: "Y," "N," and "UNKNOWN." If the interest lies in building a predictive model to predict when attribute A will have the value "Y." Samples having the value "Y" for attribute A are therefore categorized as target group **210**. The remaining attributes B, C and D for the samples are then compared, with the attribute values within target group **210** being compared to the attribute values for entire sample **204**. Attributes with the largest difference between the target group **210** and the entire sample **204** are selected.

Suppose attribute B, for instance, has five possible values within sample **204**, but target group **210** only includes two of those values for attribute B among constituent

samples. Similarly attribute C also has five possible values, four of which are exhibited by members of target group **210**. Attribute D has ten different values within sample **204**, but only three of those values are found for attribute D within target group **210**. In this case, attribute B has a larger relative difference between target group **204** and sample **204** than attribute C, since it has less overlap in attribute values. Attribute D exhibits a greater difference or divergence between samples having the desired attribute and the whole sample population than either of attributes B or C. Thus, attribute D would be selected in preference to attributes B or C in generating a predictive model, and attribute B would be chosen over attribute C.

The example described above utilizes four attributes and a relatively small number of possible values for each attribute. In practice, however, the process described may be applied to samples each having 500 or more attributes, with as many possible attribute values as there are samples for some attributes. Known statistical parameters such as entropy:

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

and/or chi-square may be utilized to evaluate the attributes to determine which exhibit the greatest difference.

In the exemplary embodiment contemplated, a predetermined number of attributes (e.g., ten) which exhibit the greatest difference with respect to attribute values, other than thsoe for the desired attribute or group of attributes, for the target group **210** versus the entire sample **204** is

selected. However, any arbitrarily sized subset of attributes selected based on entropy may be employed. Additionally, the number of attributes selected for employment in generating a predictive model need not be a fixed size, but may instead be a percentage of the total number of attributes, or only those attributes in which the values for sample instances having the desired characteristic exhibit a threshold amount of difference from the corresponding attribute values for the entire sample population.

Once the attributes which are to be employed in generating a predictive model have been selected as described above, module **202** identifies the selected attributes (e.g., by a list of attribute names) for model generator **214**. Model generator **214** may then generate a predictive model based on sample **204** utilizing the selected attributes in accordance with techniques known to those skilled in the art.

With reference now to **Figure 3**, a high level flow chart for a process of selecting attributes of a sample for generating a predictive model in accordance with a preferred embodiment of the present invention is depicted. The process begins at step **302**, which depicts a model build being initiated. A data set, from which a sample population may be drawn including at least one sample having a desired attribute, should be available for building the desired predictive model. If less than the entire data set is employed in generating the predictive model, the resulting predictive model may then be applied to the remaining samples in the data set. The desired attribute(s) for which

the predictive model is generated need not have only two possible values (e.g., gender), but may be a relative measure such as a value exceeding a predetermined threshold (e.g., monthly usage of a service).

The process first passes to step **304**, which illustrates grouping the elements of the sample population based on the values of the attribute(s) to be the subject of prediction, identifying a target group of samples. The process then passes to step **306**, which depicts selecting an attribute and determining a relative difference or divergence in the attribute values for target group samples versus the whole sample population. A relative difference (e.g., ratio or percentage) should be determined since comparison of absolute differences may not be meaningful.

The process then passes to step **308**, which illustrates a determination of whether all attributes available for the sample population, other than those for which the predictive model is being built, have been considered. If all attributes for the sample population have not been considered, the process returns to step **306** to select another attribute for analysis and repeat the process of steps **306** with the newly selected attribute.

Once all attributes for the sample population have been analyzed, the process proceeds from step **308** to step **310**, which depicts selecting n attributes exhibiting the largest relative differences for samples having the desired attributes as compared to all samples within the sample population. A sort or ranking of the attributes by such relative difference may be useful in this step. The number

n of attributes selected may be any arbitrarily set number or, as described above, may be a predetermined percentage of the attributes or attributes exhibiting a relative difference between samples which exceeds a predetermined threshold.

The process next passes to step **312**, which illustrates building a model for the desired attribute and the sample population utilizing the selected attributes. Various known techniques may be employed for this purpose. The process passes then to step **314**, which depicts applying the predictive model generated to a data set. Finally, the process passes to step **316**, which illustrates the process becoming idle until another model build is undertaken.

The present invention allows data collections have large numbers of potentially irrelevant or meaningless attributes for each sample to be employed in building an accurate predictive model. Efficiency in generating the predictive model is improved by reducing the number of attributes which are considered during the model build. This requires both less time and less computational resources to generate the predictive model. Accuracy of the resulting predictive model is also improved. Attributes which might skew the sample population but have no relation to the desired characteristic--or less relation to the desired attribute than other attributes--are eliminated from consideration in building the predictive model.

It is important to note that while the present invention has been described in the context of a fully functional data processing system and/or network, those

skilled in the art will appreciate that the mechanism of the present invention is capable of being distributed in the form of a computer usable medium of instructions in a variety of forms, and that the present invention applies equally regardless of the particular type of signal bearing medium used to actually carry out the distribution. Examples of computer usable mediums include: nonvolatile, hard-coded type mediums such as read only memories (ROMs) or erasable, electrically programmable read only memories (EEPROMs), recordable type mediums such as floppy disks, hard disk drives and CD-ROMs, and transmission type mediums such as digital and analog communication links.

While the invention has been particularly shown and described with reference to a preferred embodiment, it will be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention.